DOCUMENT RESUME

ED 300 393 TM 012 338

AUTHOR Howards, Melvin
TITLE Testing: Illusions of Measurement.
PUB DATE 87
NOTE 32p.
PUB TYPE Viewpoints (120) -- Information Analyses (070)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Accountability; Diagnostic Tests; Elementary Secondary Education; History; Literature Reviews; *Standardized Tests; Teacher Participation; Teacher Responsibility; Test Construction; Testing; *Testing Problems
IDENTIFIERS Criticism; Nation at Risk (A)

ABSTRACT

A critical position paper on research on many types of tests and the history of the testing movement is presented. The report, "A Nation at Risk" receives particular criticism. Although the accountability of teachers, schools, and educational administrators is seen as necessary, it is argued that a means of insuring accountability has not been discovered. Testing often lacks objectivity, fairness, and appropriateness and is rarely apolitical. Serious flaws in standardized tests are identified. Research shows that tests given as part of a battery (as in the diagnosis of suspected learning disabled students) are incompatible and incomparable--they do not measure the same things in the same way. In most standardized tests, the fuzzy and indiscriminate concept of average is given free play. That problems of transfer (i.e., the relationship between tasks on tests and actual performance) and that cultural differences are not accounted for in almost all standardized tests are also discussed. Teachers need much more power to influence and guide the development of school curricula, choice of textbooks, tests, and other materials if they are to be held accountable. The need for informal, individual assessment techniques is suggested; this is particularly necessary in diagnosing reading and language problems. A 52-item list of references is included. (TJH)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED300393

Testing: Illusions of Measurement

Melvin Howards, Ph.D.

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Melvin Howards

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Northeastern University
Boston, MA

BEST COPY AVAILABLE

02 FE 018

TM012338

Introduction

Never before in history have so many been tested and measured in so many different ways each year. Conservative estimates suggest that about 2000 different tests are administered to more than 200 million people annually. There are standardized tests in every conceivable aspect of human growth and development. There are CRT's (Criterion-referenced tests), many informal instruments, new tests such as the DRP (degrees of reading power), interest inventories, psychological and projective tests, including measures of personality, and others. Not only do we have a test for just about everything, we do what Gregory Anrig, head of ETS, said in a newspaper story (reported in the Boston Globe, June 30, 1985:) "There are too many tests. . . It's gotten to the point where, if it moves, we test it." Yet, ETS does more testing than any other agency, and they recently received the commission to develop the National Assessment of Educational Progress (NAEP) tests for the future. These standardized tests are given to 9, 13, and 17 year olds all across the country every few years. His quote is supposed to create the illusion of honesty.

Perhaps a definition of illusion will guide us through the research into testing in a more sharply focused manner. The OED (Compact edition, 1984) says that illusion is:

> "...the action of deceiving...a misleading image presented to the vision...something that deceives or misleads intellectually...a pattern capable of reversible perspective..." (p. 600)

The definition accurately characterizes much of what has

been happening in the testing movement since its unexpected inception with Binet in 1904. His development of some 50 tasks to identify which students were slow learners was not intended by him, nor ever claimed by him, to be a measure of intelligence. In 1905 he said:

> "The scale, properly speaking, does not permit the measure of intelligence, because intellectual qualities cannot be measured as linear surfaces are measured." (p. 40)

This honesty and unpretentiousness was ignored by the men who Americanized the scales. Goddard brought the scales to the U.S., and shortly thereafter Terman (of Stanford University) revised and expanded the instrument in 1916 and 1937. About the same time, Yerkes of Harvard, Brigham, along with Terman and Goddard, launched the testing movement with the Army Alpha and Beta tests during W.W. I.

These men held very similar views about intelligence and about measurement. They were hereditarians who believed intelligence (and almost everything else) was inherited and that the winners in this genetic roulette were white males of Northern European stock, and the losers were minorities. Terman's small, unrepresentative study of lower class youths in California in 1916 speaks to this bluntly:

> "They represent the level of intelligence which is very, very common among Spanish-Indian and Mexican families of the Southwest and also among negroes. Their dullness seems to be racial. . . Children of this group should be segregated in special classes and be given instruction which is concrete and practical. They cannot master abstractions." (pp. 91-92)

Terman recanted in 1937, but much of the damage had been done and

myths and misconceptions like this linger. The most significant and impressive development was the massive testing of army recruits in W.W. I with the Alpha and Beta tests which Yerkes, and his colleagues devised to measure the 'intelligence' of 1.75 million men. Yerkes was an hereditarian. Gould (1980) details the nature of the tests, the often haphazard administration of them, and the general insensitivity to those recruits who were either not English speakers not well educated, or blacks. In spite of the dubious and often invalid results, Yerkes reported in his 1921 monograph:

> "All officers without exception agree that the negro lacks initiative, displays little or no leadership, and cannot accept responsibility. . . All officers seem to agree that the negro is a cheerful, willing soldier, naturally subservient. (p. 742)

His general conclusion was that darker skinned people were intellectually inferior to the white, Northern European males, while his colleague Goddard (1919) said:

> "Democracy means that the people rule by selecting the wisest, most intelligent and most human to tell them what to do to be happy." (p. 237)

He was referring, in part, to the immigrants who were in the Army and whose culture and language, habits, etc., he considered inferior. The Alpha and Beta tests provided many of these testing pioneers with data to support their hereditarian biases. These notions did not die; they can be found today in the work of Eysenck, Jensen, Hernstein, Shockley, among others.

In addition to this pervasive and distorting hereditarian bias in early test making and test scoring, was the common belief that multiplex behaviors such as intelligence could

somehow be reified, transformed into a single, unitary entity. The Binet scales as modified and expanded by Terman were called intelligence without the benefit then or now of a clear, universally accepted definition of intelligence. In short, intelligence is what they called it. Current research into intelligence measurement by Kaufman, Sternberg and Gardner speaks of many kinds of intelligence; not the reified, single, unitary entity of the past. This practice of reification is verbally convenient but conceptually very misleading. Another of those multiplex human behaviors that we think we can control and limit by giving it a name is perception, even though it is not a single, unitary entity. Like intelligence, memory, reading, it is comprised of many elements in constant interaction on several levels with each other.

This naming fallacy, this reification, has created a major illusion in the field of measurement, which has blended well with the hereditarian biases of the men who launched the testing movement. It prepared the field for the growth and development of numerous other illusions. Brigham, a disciple of Yerkes of Harvard, became secretary of the CEEB which developed the SAT, recanted his early hereditarian biases in 1930, saying:

> "Most psychologists working in the test field have been guilty of a naming fallacy which easily enables them to slide mysteriously from the score in the test to the hypothetical faculty suggested by the name given to the test. Thus, they speak of sensory discrimination, perception, memory, intelligence and the like. . . (p. 159)

He also admitted that the Army tests really measured

familiarity with the English language and with American culture rather than innate intelligence. This recantation was too little too late to dilute the well established racial and ethnic hereditarian biases.

Other deleterious and pervasive myths and illusions are: ordinal scales can measure non-ordinal qualities such as intelligence, perception, memory, etc.; that tests found to be invalid and/or unreliable may be used to measure performance and progress; that tests with small and/or unrepresentative standardization populations can be used to rank order students; that standardized tests are not culturally biased against minority groups, especially racial and ethnic minorities; that several tests purporting to measure the same thing, e.g., intelligence, reading comprehension, vocabulary, etc., are comparable; that the concept of average is useful in diagnosis of learning problems; that test scores provide a sound basis for prediction of future performance, behavior, and that there is good transfer from test performance to learning situations, and that the normal distribution curve is something other than a mathematical convenience.

After its birth and early development, the testing movement became a business, an industry in fact. Its major product, perhaps its only one, is illusion: the illusion of precision, of comparability of persons, of expectations, of progress or its lack, of ranking, of prediction, of objectivity and truth. All of the illusions mentioned above will be illustrated in detail when we look at what research says about many of the

most frequently used tests. Before we do that, a few words about problems of administering many different tests to any individual, about 'average', about transfer and, lastly, about culture.

Research has shown very clearly that tests often given as part of a battery as in the diagnosis of suspected learning disabled students, for example, are incompatible and incomparable: that is, they do not measure the same things in the same way. This means that test administrators, psychologists and other specialists who evaluate these test scores must choose which scores on which tests to accept and to interpret.

This selectivity is subjective and it is not based on any verifiable principle or theory. The selection often appears to be the result of a predisposition and prejudgment of a 'suspected' student, and also the result of expediency. Naturally, one cannot exclude from this equation the power of money.

Another factor is that in most standardized tests the fuzzy and indiscriminate concept of average is given free play. Average scores are misleading. For example, five persons earning a 100 IQ on the WAIS or Binet will be found to have earned that score in very different ways from each other. Their test profiles will show quite a variety of high and low scores, but the average of all the subtest scores is calculated to be 100. That number tends to be reified as 'average', which is reified to mean normal, doing what he/she is supposed

to be doing at a particular time in life. But, clearly, it is a very poor discriminator of actual strengths and weaknesses of any individual in any particular area.

Another crucial factor that is rarely discussed in testing is that of transfer. In this context I mean the relationship, or lack thereof, between tasks on tests and actual performance. If a student scores well on a vocabulary subtest, does it mean that the student will be able to understand those words in a variety of contexts? Will the student understand the multiple meanings of some of these words? Will the student understand metaphoric or symbolic meanings? Will the student incorporate the new words on new meanings of older words in his/her speaking learning?

A final example to clarify my point. In many reading readiness programs children are taught to discriminate among various geometric forms. Those students who can do this are said to be more likely to be able to learn to read. The misleading analogy underlying this is: reading means dealing with words, words are composed of letters and letters are shapes, therefore, if children can be taught how to discriminate among geometric shapes, it may assist them in learning how to deal with letter shapes. Not only does research contradict this notion, but simple logic says that discriminating among neutral, non-affective geometric shapes and forms is qualitatively different cognitively and affectively from discriminating among letters in meaningful words.

This analogy applies to other analogies of transfer: if

you can do a task on an artificial, often non-contextual, stressful test, you can do it in connected, meaningful material. Patent nonsense!

And lastly, cultural differences are not accounted for in almost all standardized testing. If culture is not accounted for, we omit the pervasive and universal power of a specific language, certain customs, values, beliefs, logic, feelings, traditions, attitudes, perceptions of time, space, beauty, etc. In sum, if we omit culture, we deny the individuals their identity, their unique blend of qualities, and if we do this, we do not measure them at all. The results of the administration (often quite sloppy) of the Army Alpha and Beta tests and the interpretation of the culturally biased items, should have warned us of the dangers of omitting culture, but it has not. We continue to propagate the familiar illusion of what is right and normal as that which has been devised by and interpreted typically by white middle class Northern European for others of similar background.

What does the research on modern testing say?

The Research

Haggard reports in Schell (1981) on the 1973 Woodcock Reading Mastery Tests

> "The most troubling and disappointing aspect of the tests' construction is the absence of any explanation as to why these particular subtests were decided upon and why some unusual formats, e.g., analogy in word comprehension, were used. . . The validity of the subtest is therefore quite uncertain." (59) the usefulness of the entire battery as a before and after measure of reading achievement for secondary

students seems severely limited." (60) The analogy format: "...the concern is that such items measure something besides word knowledge, and that there is no way to tell when they are measuring what...the word comprehension test does not measure a clearly defined word comprehension skill; results may be contaminated by an unknown degree by variations in such abilities as abstract reasoning, general intelligence, and learning speed." (62)

Gillet and Temple (1986) add:

". . . on subtests like word identification or passage comprehension does not tell the teacher what the student can or cannot do. The examiner still does not know what letters the student can identify or what phonics skills have been mastered without detailed analysis, which is not discussed in the manual." (pp. 329-330)

Salvia and Ysseldyke (1985):

". . . some subtests must be used with a degree of caution." (360)

Johnston in Pearson (1984) adds that:

". . . certain reading tests, such as the Woodcock Reading Mastery Test (1973) or the Stanford Diagnostic Reading Test which claim to be both norm-referenced and criterion-referenced. . . should be looked upon with considerable suspicion. In their attempts to serve two masters, they may serve neither adequately." (159)

Gillet and Temple (1986) compared the popular Peabody Individual Achievement Test (PIAT) with the even more popular Wide Range Achievement Test (WRAT):

"Like other achievement tests, the results they yield are solely quantitative, and they have little, if any diagnostic value. . ." (336)

The same authors also compared the WRAT with the Slosson Oral Reading Test (SORT) and concluded that since both tests purport to measure reading by using lists of single words, neither measures "real reading at all". (p. 336)

Mehrens and Lehman (1987) found that the PIAT was more a

measure of speed than power as is claimed, and they also found low subtest reliabilities. Hansen in Schell (1981) points out that in the PIAT:

> ". . . the tester must judge the students' pronunciation; acculturation factors affect the outcome. . . the validity of the test would be questioned by most educators." (p. 49)

Except those who continue to use the test and to misinterpret it. Hansen concludes her remarks on the PIAT:

> "Despite impressive standardization procedures, these subtests can't be recommended very highly." (p. 52)

To complete these samples of criticism, a word from Salvia and Ysseldyke (1985) on the WRAT:

> "The test has limited behavior sampling, inadequate standardization and absence of reliability and validity data." (p. 329)

What remains?

Criterion-referenced tests became popular in the 1970's and remain somewhat popular even now. They did serve a purpose by pointing out that norms and grade level equivalents did not help classroom teachers or specialists know exactly why students got some items correct and some incorrect. Some professionals felt the need to know in more detail how students were reading and responding. Criterion-referenced tests have their shortcomings, perhaps the most significant of which is that the CRT's test minute sub-skills with two or three questions which is much too small a sampling of normal reading activity and behavior. Also, the CRT's have broken the reading process down into so many discrete subskills that such a measure produces

a distorted and unrealistic assessment of real reading.

Gillet and Temple (1986):

> ". . . Like other tests, criterion-referenced tests frequently include tasks that are easily measured but not very closely related to reading." (p. 343)
> "It is difficult, if not impossible, to isolate skills in actual reading because meaningful reading requires the use of many skills and processes simultaneously, and partitioning overlooks this important factor." (p. 346)

To highlight these remarks, when the Boston Public Schools moved to criterion-referenced testing in the early and mid-1970's, their handbook describing the program listed 111 separate reading skills! Impossible, implausible, and contrary to all we know about the reading process, about learning, and about people.

The Developmental Test of Visual Perception (DTVP) by Marianne Frostig, et. al., has, by the author's own admission, some very critical shortcomings: a middle class sample taken from a limited geographical area, with little minority representation. Salvia and Ysseldyke (1985) report that:

> ". . . individual subtests certainly raise serious questions about their use in differential diagnosis, the very procedure Frostig recommends. . . The validity data reported in the manual for the DTVP do not support the authors' contention that the test measures five operationally defined perceptual skills. . . Individual subtests of the DTVP lack the necessary reliability and validity to be used in diagnostic prescriptive teaching. We simply cannot put a great deal of faith in the accuracy (freedom from error) of the scores a child earns on the DTVP subtests." (275-6)
> "What the majority of the research <u>has</u> shown is that most perceptual-motor tests are unreliable. We do not know what they measure, because they do not measure anything consistently. . . And for the most part they are neither theoretically nor psychometrically sound . . ." (267)

Hammill and Wiederholt (1973) and Larsen and Hammill (1974)

did reviews of the massive correlational research on several aspects of visual perception:

> "It was concluded that neither the total score nor the sub-test scores of the DTVP relate meaningfully to reading."

ITPA (Illinois Test of Psycholinguistic Abilities, 1968).

This test is still used by many learning disability specialists, speech pathologists, school psychologists, and others who are responsible for determining who does or does not have a learning disability. The use of this instrument flies in the face of a long history of extremely critical reviews of its conspicuous weaknesses. A sampling of this very persuasive criticism follows.

Salvia & Ysseldyke (1985):

> "The ITPA was standardized on a restricted group of children. . . although the test was designed 'for use with children encountering learning difficulties'. . . children who were having problems were systematically excluded from the normative sample. In fact, the children included in the norm groups were 'only those children demonstrating average intellectual functioning, average school achievement, average characteristics of personal-social functioning, sensory motor integrity, and coming from predominantly English speaking families' . . . School age children were selected by sampling only schools that administrators judged to be predominantly middle-class." (258)
> ". . . the reliability of the ITPA is far from adequate. . . Validity data presented in the ITPA manual is unconvincing." (259)
> ". . . the subtests of the ITPA rarely assess the behaviors that the authors claim they assess." (275)

Smith and Marx (1972) noted that the ITPA correlated fairly well with the WISC, r=.63, then query:

> "Are we then to consider psycholinguistic abilities as facets of intelligence or of language development? What is the ITPA adding to our knowledge about the child if it is, in effect, largely a measure of verbal intelligence,

> other than perhaps new names for old subtests of intelligence?" (pp. 357-362)

They go on to say that some of the sub-tests may not measure what they say they measure:

> "Auditory Reception may demand organizing rather than receptive processes; visual reception involves both memory and conceptual thinking and may be verbal as well as visual. . ." And they conclude: "The groupings suggested by the authors for the use of the profile for diagnosis of areas of deficit are hardly supported by the relationships among the tests indicated by the factor analysis." (p. 359)

Carroll's review of the ITPA in the Buros' Seventh Mental Measurements Yearbook (1972) agrees with Smith and Marx but goes even further in questioning the limited standardization population, (about 1600 students in and around Decatur, Illinois), the excessive testing time, the complexity of administering the battery, and even whether the test is a measure of psycholinguistic abilities since only about half of the tests involve the use of language. He also found that there was no pattern of high or low scores that are characteristic of children with reading disabilities.

Spache (1976) concludes his analysis and discussion of the ITPA:

> "Most of the subtests are affected by socioeconomic status, as we would expect for an instrument standardized on a population drawn from various levels of the middle class."

Newcomer and Hamill (1975) found that the diagnostic validity of the ITPA in discriminating between groups of good and poor readers was completely negative for all subtests when intelligence was controlled.

Clearly, many researchers find the ITPA to be all but useless for diagnostic work, particularly for populations different from the very small and relatively homogeneous population used to standardize this collection of poorly defined subtests.

Just as there is a major problem with reifying intelligence and then attempting to measure it with a variety of tasks, so is there a problem with the reification of perception and trying to measure it. It is not clear to most observers what perception is. Some of the ITPA subtests suggest they are tapping into some kind of perception, as do the Beery Visual Motor Intelligence Test (VMI), the Bender Visual Motor Gestalt Test (BGVMGT) among others. Research reveals the problems. Guthrie, Goldberg and Finucci (1972) studied the relationship between visual perception of geometric forms and the ability to read and found:

> "Tests of the two show low correlations, and many cognitive operations--such as perception of redundancy and pronunciation, which are required for word recognition--are absent in abstract pattern perception."

Guthrie & Seifert (1978) add:

> ". . . visual perception is a process required for reading or at least word recognition. And visual perception is also a process required to remember a sequence of geometric forms. . . However, the two sets of processes--those needed to perceive words and those needed to perceive geometric forms--are different."

Porter and Binder (1981) studied visual motor tests and concluded:

"The Beery Visual Motor Intelligence (VMI) and the

> Bender Gestalt Test (BGT) are theoretically related. Practically, however, the VMI and the BGT display only a limited practical inter-test reliability. . . although sharing some common variance, also measure different constructs in reference to the assessment of the child's level of visual-motor development."

This speaks to the well documented incomparability of many of the tests used in diagnosing reading, learning, and language problems. These two visual-motor tests actually measure different aspects and qualities of visual-motor development and skill.

A final and devastating word on the Bender from Salvia and Ysseldyke (1985):

> "There is no evidence about the extent to which the test assesses visual-motor perception; the copying of nine designs is believed to be a measure of visual perception because some experts say it is." (270)
> "The BVMGT is, quite simply, a measure of skill development in copying geometric designs. It is not designed as a measure of intelligence, predictor of achievement, or measure of emotional disturbance or minimal brain dysfunction." (272)

Nowhere in the testing movement, past or present, can we find more abuse, misuse and extreme selectivity in the interpretation of results than in the diagnosis of children suspected of having ill-defined, vague, and encompassing learning disabilities. Elsewhere I have dealt with some of the key issues related to learning disabilities in considerable detail: the imprecise definitions of it, the inability, therefore, to determine the incidence of learning disabilities, the misuse of much of the typical test battery for diagnosing learning disabilities, and finally, the often ineffective remedial or therapeutic programs to alleviate the learning disabilities. Some of the leading researchers in testing and measurement say:

Ysseldyke, et. al., (1981) sum up the situation:

> "We need a new perspective on assessment; administering standardized tests to confirm teacher observation is wasteful, administering standardized tests to plan instructional programs is ineffective, and administering standardized tests to evaluate educational progress is failing."

Campbell, (1979) says:

> "In the past, diagnostic workups have often been inadequate and poorly focused relying too heavily on test scores without thorough evaluation of the child in his or her environmental context. The identification of skill deficit masquerading as diagnosis cannot lead to reasoned clinical decisions about appropriate treatment strategies." (My italics)

The last sentence is absolutely critical but almost always overlooked.

It is striking how many different researchers have reported similar findings on so many commonly used tests, yet practice has not changed:

Coles, (1978):

> "There is growing disillusionment with the practice of diagnosing and remediating LD children on the basis of the standard LD battery which includes the use of standardized tests in the assessment of the child's intelligence, memory, language, auditory discrimination, psychomotor skills, and neurological integrity."

Arter and Jenkins, (1978):

> "The reasons for this disillusionment are many, including the questionable statistical characteristics of many commonly used instruments. . . the present means of assessing specific strengths and weaknesses are either inadequate or unimportant in remediating school-related difficulties."

A popular test that has earned the same kind of criticism as those above is the Peabody Picture Vocabulary Test (PPVT). In its original unrepresentative standardization, the Peabody

included about 4000 white, middle class students from the Nashville, Tenn. area. The test was revised in 1981 and now has a more representative population and its reliability has improved but, "There are no data in the PPVT-R manual on the validity of the test," Salvia and Ysseldyke (1985).

Memory tests are normally included in the typical learning disability battery as well as in other diagnostic procedures. Ceci, et. al., (1981) found:

> "Numerous correlational studies of well-known memory tests (e.g., Detroit, G-F-W, WISC-R span, ITPA Visual sequential memory, MFD, Benton Visual Retention, Wepman) have shown very low intercorrelations between them suggesting that the cognitive skills being tapped by these instruments are not identical."

After all of this testing and measurement, what are the effects, the outcomes? Algozzine and Ysseldyke (1986) answer:

> ". . . observational studies of placement team meetings suggest there is little relationship between data presented as evidence of problems and actual decisions presented at the meeting. . . and classroom observations indicate that little changes in the instructional ecology of students as a result of the diagnostic process." (396)

Even though many of the widely used tests are very seriously flawed in several major respects, the testing juggernaut moves on, crushing logic, reason, no matter what research says.

What has been said thus far about intelligence tests, memory tests, perceptual tests, etc., is also true of reading tests, perhaps the most widely administered of all tests. Many federal and state programs dole out untold millions annually to support various remedial reading programs such as Chapter I, yet the identification of those who need special help is

derived primarily from the results of standardized reading tests.

Karlsen in Schell (1981) paints the picture vividly:

> ". . . since there is no professional consensus as to the various skills in reading, no two tests include identical subtests. In fact, several tests reviewed herein for the same reading level were more different than alike." (p. 5)

Farr and Carey (1986) summarize very well how incompatible the most commonly used standard reading tests are. They compared the Metropolitan Achievement Test (MAT), the California Test of Basic Skills (CTBS), Iowa Test of Basic Skills (ITBS), California Achievement Test (CAT), and Stanford Achievement Test (SAT): The MAT measures sight words and so does the SAT. The CTBS and the others do not; MAT measures visual discrimination of shapes, letters or words and so does the CAT, but none of the others does; CTBS, CAT, and SAT measure syllabication, the others do not; recognition of root words is measured only by the CTBS; rhyming words is included in the CTBS and ITBS, and so it goes. The authors comment:

> "The lack of consistency across the tests does not stop with these differences. Often, there is variation as to when certain skills are tested." (p. 77)

And they report what most of us already knew but have done nothing about:

> "There is no question that the administration and use or misuse, of reading tests in 1984 has increased significantly over the use of tests fifteen years ago. . ." (p. 6)

Valencia and Pearson (1986):

> ". . . currently available tests do not reflect the

> advances in reading theory, practice or research. They are based on one view of the reading process, while the best reading instruction is based on an alternative and contradictory point of view. As the result, there is a serious discrepancy between what we know and what we measure."

They conclude that,

> ". . . the picture of a nation of schools, teachers and students engulfed by tests is complete. . . the influence of testing is greater now than at any time in the history of schooling." (p. 2)

Clearly, very few professionals are reading this research, or else they ignore it. Some educators say, when faced with this mass of critical research, yes, but . . . Yes, but if you know how to select and interpret the particular results on particular tests, they can be helpful in diagnosing and correcting reading and learning problems. The key words are select and interpret. If one is willing to omit the data on lack of validity and reliability on many of the tests already cited, and if one does not care much about the size and representativeness of the standardization population for norms, and if one is willing to define reading, for example, as a large number of isolated, fragmented skills, then of course professionals will continue to perpetuate the illusions of measurement. In light of this much research challenging the usefulness and validity of many commonly used tests, it is much more than ironic that there is more testing now than ever before.

In the national study Becoming a Nation of Readers (1985) we hear:

> "Performance on standardized tests can be influenced by strategies that are not used in normal reading. . .

> most tests do not permit skilled readers to use strategies that are important in normal reading. . . Standardized tests give an impoverished picture of reading competence. (pp. 98-99)

We have seen that the change to CRT's was a minor improvement in testing, but that it too was burdened by more disadvantages than advantages. More recently, Degrees of Reading Power (DRP) tests have been used in many school systems. The DRP attempts to measure reading and language competence in meaningful contexts, carefully graded by readability, and using the cloze procedure. But the cloze procedure contains serious flaws that may actually confound the results.

Cloze procedure requires the student to choose the right word to fit into the randomly omitted word, e.g., every nth word. More often than not, however, more than one word may fit into such a blank and make good sense.

Gillet and Temple (1986) warn that, ". . . there is considerable ambiguity in the process." (p. 193)

Farr and Carey (1986):

> ". . . there exists a common measurement factor (in terms of shared variance) between cloze and conventional multiple choice tests of comprehension. . . Since we know there are problems with multiple choice tests, do we want to employ a device which correlates very highly with these tests? . . . to make decisions based on the use of cloze only is to make less than informed choice." (p. 37)

Miscellaneous Tests

Let me conclude this survey of testing with a sweep of several different types of tests and what some researchers have to say about them.

There is no reliability data provided for the Sucher-Allred Placement Inventory and it only provides a rough estimate but does not measure reading achievement. The SORT (Slosson Oral Reading Test) is not an oral test in spite of its name and no attempt is made to measure knowledge of word meanings nor of comprehension. Gillet and Temple (1986) conclude that, ". . . the reading level this test yields is largely useless." (p. 328) Incidentally, the same could be said of any oral reading test that depends primarily on pronunciation without any measure of comprehension of words or ideas in sensible context.

Ysseldyke and Salvia (1985) say of the Durrell Analysis of Reading Difficulty (DARD, 1980):

> "The norm-referenced use of the test is hindered by inadequate standardization, absence of a description of the norm group, limited data on reliability, and limited validity." (335)

Why do Durrell and others who have made big names for themselves over many years in the field of reading and of testing, omit such essential data as this? These people surely have not forgotten that such data is necessary for any test. Do they believe that their reputations justify significant omissions like this: Are we to take them at their word? If so, then we have to deal not only with the various illusions of measurement these instruments have shown us, but we must also suspend our disbelief about their motives because of the illusion of celebrity.

Salvia and Ysseldyke (1985) referring to the Otis-Lennon

School Ability Test (1979) say:

> ". . . there are some serious questions about the adequacy of standardization of the scale. Data regarding validity are inadequate." (207)

And then they report something many professionals may not know or recognize:

> ". . . most group intelligence tests, while standardized on large numbers of students, often are not standardized on representative populations. Most are standardized on districts, not on individual students. . . on volunteer samples, if one drops out it is replaced by another district which may not be properly comparable and may bias the standardization." (pp. 211-12)

The same authors take aim at another widely used diagnostic test:

> "The Gates-McKillop-Horowitz Reading Diagnostic Tests are a widely used diagnostic instrument, in spite of significant limitations. The manual provides normative tables without including a description of the population on whom the test was standardized. The many scores obtained on the tests are subject to misinterpretation. Evidence for reliability and validity is unsatisfactory." (p. 351)

Space requires an end to this survey, but a very important fact about all standardized testing is the normal distribution curve and needs to be mentioned. It is one of the most pervasive of all the illusions of measurement. Lewontin, Rose and Kamin (1984) sum it up precisely:

> ". . . the famous normal distribution--is arbitrary, but its power is such that psychometrists chop and change their scales till they meet this criterion. . . Psychometry, above all, is a tool of conformist society. . . mainly concerned to match them against others and to attempt to adjust them to conformity. . ." (p. 93)

Our notions and judgments about students being average, above or below average, are related to this misleading, artificial

mathematical concept. The illusion of objectivity is thus created.

Conclusions and Recommendations

In the face of this overwhelmingly critical research of many types of tests, the national study of American education that helped launch the much talked about 'reform' movement, A Nation at Risk, blithely calls for more testing! It is as if the writers of the report and some well placed educational leaders, a) do not read this research, b) do not believe it, c) do not care about critical research, d) have a different agenda. They all justify more testing by saying that it is the only way, or at least the best way they can think of, to hold schools, teachers, administrators, everyone (except themselves, it seems) accountable.

Accountability is certainly needed, but we have not been able to decide how to do it farly, objectively, appropriately and non-politically. Teachers need much more power to influence and guide the development of school curricula, choice of textbooks, tests, and other materials if they are to be held accountable. Can they be held fully accountable for the failures of the school administration, the publishers, the curriculum makers, specialists, experts, politicians, parents, and community at large?

Obviously, many standardized tests are seriously flawed in major respects and should not be used as they are now, but we need some reliable way to assess student performance and

progress. What might be done? One alternative is some form of informal, individual assessment. This is especially helpful in diagnosing reading and language performance. Farr and Carey (1986) agree:

> ". . . informal testing holds the greatest promise for increasing validity and reliability of information for instructional planning." (p. 175)

Others also see the value of such assessment, particularly in reading. But most people point out that one cannot do informal, individual measurement due to class size, lack of objectivity of the informal assessments, reliability, etc. None of these is valid except for class size, but that, too, can be managed without too much effort. Teachers need to be taught how to administer representative samplings of material and to gather the essential information on how the student performs. What are his/her specific weaknesses and strengths when dealing with meaningful, connected text? This is an area that needs more study and development; we do not know all of its potential yet, nor all of its problems. We do know most of the problems of current testing, and it is time to dissolve many of the current illusions of measurement.

Let me close with remarks from O.K. Buros (1972):

> "At least half of the tests currently on the market should never have been published. Exaggerated, false or unsubstantiated claims are the rule rather than the exception."

REFERENCES

Algozzine, Bob and Ysseldyke, James E. (1986). The Future of the LD Field: Screening and Diagnosis. Journal of Learning Disabilities, 19, 7, 394-398.

Ames, Louise Bates. (1983, Jan.). Learning Disability: Truth or Trap? Journal of Learning Disabilities, 19-20.

Arter, J. A. and Jenkins, J. R. (1978). Differential Diagnostic-Prescriptive Teaching: A Critical Appraisal. Technical Report no. 80. Center for the Study of Reading: University of Illinois.

Becoming a Nation of Readers. (1985). The Report of the Commision on Reading. The National Academy of Education. The National Institute of Education. The Center for the Study of Reading. Prepared by Richard C. Anderson, Elfrieda H. Hiebert, Judith A. Scott, Ian A. G. Wilkinson, Washington, D. C.

Brigham, C. C. (1930). Intelligence tests of immigrant groups, Pyschological Review, 37, 158-165.

_____. (1923). A Study of American Intelligence. Princeton, NJ: Princeton University Press.

Buros, O. K. (1972). The Seventh Mental Measurements Yearbook. (Ed.) Highland Park, NJ: Gryphon Press.

Campbell, Susan B. PL 94-142: Diagnostic Problems. Journal of Learning Disabilities, October, 1979, 511-513.

Carroll, John B. in O. K. Buros. (1972). Seventh Mental Measurement Yearbook. (Ed.) Highland Park, NJ: Gryphon Press.

Ceci, Stephen, Ringstrom, Maureen and Eglea, Stephen. (1980, March). Do Language-Learning Disabled Children (L/LD's) have Impaired Memories? In search of underlying processes. Journal of Learning Disabilities, p. 160.

Colarussso, R. P., Hartig, J. and Martin, H. (1975). Specific visual perceptual skills as long-term predictors of academic success. Journal of Learning Disabilities, 8, 651-655.

Coles, G. (1978). The Standard LD Battery. Harvard Educational Review, 48, 313-345.

Salmon-Cox, L. (1981). Teachers and standardized achievement tests: What's really happening? Phi Delta Kappan, 62, 631-634.

Cruickshank, Wm. M. in James Tucker, Linda J. Stevens and James E. Ysseldyke. (1983, Jan.). Learning Disabilities: The Experts Speak Out. Journal of Learning Disabilities, pp. 6-14.

Edmonds, R. (1985). Characteristics of effective schools: Research and implementation. in J. Osborn, P. T. Wilson, and R. C. Anderson (Eds.) Reading Education: Foundations for a literate America, pp. 123-130. Lexington, MA, Lexington Books.

Frederiksen, N. (1984). The real test bias: Influences of testing on teaching and learning. American Psychologist, 39, 193-212.

Gillet, Jean Wallace & Temple, Charles (1986). Understanding Reading Problems. (2nd. ed.). Boston: Little, Brown and Co.

Goddard, H. H. (1919). Psychology of the normal and subnormal. N.Y.: Dodd, Mead, and Company.

Gould, Stephen Jay. (1981). The Mismeasure of Man. NY: Norton.

Guthrie, J., Goldberg, H. and Finucci, J. (1972). Independence of Abilities in Disabled Readers. Journal of Reading Behavior, 4 (2), 129-38.

Guthrie, John T. & Seifert, Mary. (1978). In Myklebust, H. (Ed.) Education for children with reading disabilities. NY: Grune & Stratton. pp. 223-255.

Hammill, D. D. and Bartel, N. E. (1975). Teaching Children with Learning and Behavior Problems. Boston, MA: Allyn and Bacon.

Hammill, D. D. and Larsen, S. C. (1974). The Effectiveness of Psycholinguistic Training. Exceptional Children, 41: 5-14.

Hammill, D. D., Wiederholt, J. L. (1973). Review of the Frostig Visual Perception Test-Related Training Program. The First Review of Special Education. vol. 1 (Edited by L. Mann and D. A. Sabatino. Phila: Journal of Special Education Press, pp. 33-48.

Johnson, Donald H. and Johnson, Leah. (1978). LD: What Research Says to the Teacher. Washington D. C.: NEA.

Kolers, P. (1975). Specificity of operations in sentence recognition. Cognitive Psychology, 7 (3): pp. 239-306.

Lewontin, R. C. Rose, Steven and Kamin, Leon J. (1984). Not in Our Genes. NY: Pantheon Books.

Lily, Steve in James Tucker, et. al. (1983), Jan.). Learning

Disabilities: The Experts Speak Out. Journal of Learning Disabilities, pp. 6-14.

Mehrens, William and Lehman, Irvin J. (1987). Using Standardized Tests in Education. (4th edition) NY: Longman.

Myers, P. I. and Hammill, D. D. (1976). Methods for Learning Disorders. (2nd edition). New York: John Wiley and Sons.

Myklebust, Helmer. (1978). Progress in Learning Disabilities. (Ed.) Vol. IV. New York: Gruner and Stratton.

A Nation at Risk. (1983). The National Commission on Excellence in Education: Washington, D. C.: U. S. Department of Education.

Newcomer, Phyllis, Hare, Betty, Hammill, D. D. and McGettigan, James. (1975, April). Construct Validity of the Illinois Test of Psycholinguistic Abilities. Journal of Learning Disabilities, pp. 220-231.

Oxford English Dictionary (Compact Edition). London: Oxford University Press (1971, 1984).

Pearson, P. David. (1984). Handbook of Reading Research. (Ed.) NY: Longman.

Porter, Gary L. and Binder, Dorothy, M. (1981, March). A pilot study of visual-motor developmental inter-test reliability: The Beery Developmental Test of Visual-Motor Integration and the Bender Visual Motor Gestalt Test. Journal of Learning Disabilities, pp. 124-127.

Ruddell, Robert B. (1985, Feb.). Knowledge and attitudes toward testing: Field Educators and legislators. The Reading Teacher, 38, 538-543.

Salvia, John and Ysseldyke, James E. (1985). Assessment in Special and Remedial Education (3rd edition). Boston: Houghton, Mifflin Company.

Samuels, S. Jay. (1973, Winter). Success and failure in learning to read: A critique of the research. Reading Research Quarterly, 8, 200-39.

Schell, Leo M. (1981). Diagnostic and Criterion-referenced Tests: Reviews and Evaluation (Ed.) Newark, Del: IRA.

Shepherd, Margaret J. (Ed.) (1983, Jan.). LD: The Experts Speak Out. Journal of Learning Disabilities, pp. 6-14.

Silver, Patricia Gillespie in Tucker, James, et. al. (1983, Jan.). LD: The Experts Speak Out. Journal of Learning Disabilities, 6-14.

Smith, J. David and Polloway, Edward A. (1979, Oct.). LD: Individual Needs or Categorical Concerns? Journal of Learning Disabilities, pp. 525-528.

Smith, Philip A. and Marx, Ronald W. (1972, June/July). Some cautions on the use of the Frostig Test: A factor analytic study. Journal for Learning Disabilities, pp. 357-362.

Spache, George D. (1976). Investigating the Issues of Reading Disabilities. Boston: Allyn and Bacon.

Terman, L. M. (1916). The Measurement of Intelligence. Boston: Houghton Mifflin.

_____ and Merrill, Maud A. (1937). Measuring Intelligence: A guide to the administration of the new revised Stanford-Binet tests of intelligence. Boston: Houghton Mifflin.

Torgeson, Joseph K. and Dice, Carl. (1980, Dec.). Characteristics of Research on LD. Journal of Learning Disabilities, pp. 531-535.

Valencia, Sheila and Pearson, P. David. (1986, Dec.). New Models for Reading Assessment. University of Illinois at Urbana, Champaign, Reading Education No. 71, Center for the Study of Reading.

Yerkes, R. M. (1921). Psychological examining in the United States Army. (Ed.) Memoirs of the National Academy of Sciences, vol. 15, 890 pp.

Ysseldyke, James E., Algozzine, B. and Epps, S. (1982). Logical and Empirical Analysis of Current Practices in Classifying Students as Handicapped. MN: Institue for Research on Learning Disabilities.

Ysseldyke, James E. and Algozzine, Bob. (1983, Jan.). LD or not LD: That's Not the Question! Journal of Learning Disabilities, 29-31.